Published online 29 October 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D301–D308
doi:10.1093/nar/gkq1069

RBPDB: a database of RNA-binding specificities

Kate B. Cook[1], Hilal Kazan[2], Khalid Zuberi[3], Quaid Morris[1,2,3,4] and Timothy R. Hughes[1,3,4,*]

[1]Department of Molecular Genetics, [2]Department of Computer Science, [3]The Terrence Donnelly Centre for Cellular and Biomolecular Research and [4]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada

Received August 15, 2010; Revised October 8, 2010; Accepted October 14, 2010

ABSTRACT

The RNA-Binding Protein DataBase (RBPDB) is a collection of experimental observations of RNA-binding sites, both *in vitro* and *in vivo*, manually curated from primary literature. To build RBPDB, we performed a literature search for experimental binding data for all RNA-binding proteins (RBPs) with known RNA-binding domains in four metazoan species (human, mouse, fly and worm). In total, RPBDB contains binding data on 272 RBPs, including 71 that have motifs in position weight matrix format, and 36 sets of sequences of *in vivo*-bound transcripts from immunoprecipitation experiments. The database is accessible by a web interface which allows browsing by domain or by organism, searching and export of records, and bulk data downloads. Users can also use RBPDB to scan sequences for RBP-binding sites. RBPDB is freely available, without registration at http://rbpdb.ccbr.utoronto.ca/.

INTRODUCTION

RNA-binding proteins (RBPs) have a fundamental role in a wide variety of cellular processes including transcription, RNA splicing and processing, localization, stability and translation (1–6). RBPs typically contain RNA-binding domains (RBDs) such as the RNA Recognition Motif (RRM) and the K homology (KH) domain, which are among the most numerous protein domains in metazoan genomes, including the human genome (7–9). Individual RBPs often have multiple RBDs that can independently bind RNA (10), and the approximately 400 annotated mammalian RBPs contain over 800 individual RBDs (11).

Knowledge of the RNA-binding activity of RBPs is critical for mapping and understanding transcriptional and post-transcriptional networks and regulatory mechanisms. Collections of DNA-binding specificities of transcription factors are available and widely used (12,13); however, to our knowledge, there is no central repository of information on the RNA-binding activities of RBPs. Here, we introduce RNA-Binding Protein DataBase (RBPDB), a database of RNA-binding experiments. A total of 1453 *in vitro* and *in vivo* experiments on 272 proteins are included, as well as 71 binding profiles in the form of position weight matrices (PWMs) and sequence logos, and 36 sets of sequences bound *in vivo* in immunoprecipitation experiments.

We anticipate that RBPDB will be of use to diverse researchers. In addition to searching for RNA-binding activities by protein, domain and experiment, RBPDB also allows users to scan RNA sequences for matches to RBP binding preferences stored in RBPDB. Additionally, the collected motifs should prove invaluable for genome-wide scans to identify *cis*-regulatory elements involved in post-transcriptional regulation via RBPs. Finally, the inclusion of *in vivo* bound transcripts provides a snapshot of enriched RBP-specific mRNA targets.

DATABASE DESIGN AND IMPLEMENTATION

Overview

RBPDB is a collection of RBPs linked to a curated database of published observations of RNA binding. The database consists of a table of proteins, linked to other proteins through orthology relationships and to one or more experiments, if experiments are found. Each protein and experiment is assigned a unique internal ID number, and proteins are linked to Ensembl, FlyBase and WormBase gene annotations and RNA-bound protein structures on PDB (14–17). Experiments are associated with a PubMed ID. Motifs, PWMs and large-scale data sets are retained as flat files that are linked to experiment and protein IDs.

*To whom correspondence should be addressed. Tel: +416 946 8260; Fax: +416 978 8287; Email: t.hughes@utoronto.ca

Protein catalog

To populate the database, we first cataloged known and predicted RBPs in human, mouse, Drosophila and *Caenorhabditis elegans* (18–26). Most proteins were selected based on the presence of known sequence-specific RBDs (Table 1), which we compiled from review papers (3,4,7,8) and from searching and scanning Pfam domain annotations (27). We retrieved protein matches to InterPro domains from UniProt and Ensembl and used the union of these two sets. Additionally, we added proteins that bind RNA through a non-canonical RBD, such as a Sterile Alpha Motif (SAM) domain or C2H2 zinc finger, based on a Gene Ontology or keyword annotation as RNA-binding in Ensembl, UniProt or NCBI. However, we did not include domains that are largely specific to ribosomal proteins (e.g. S4 domain). Moreover, some non-sequence specific, poorly characterized and/or unconventional RBDs are currently not included (e.g. dsRBD, G-patch, zinc-knuckle and zinc-ribbon) (7). Inclusion of additional domains and species is a future objective for RBPDB, and users can suggest novel domains for inclusion (see Future Directions section). We note, however, that in eukaryotes, the repertory of known and predicted RBPs is dominated by RRM and KH domains, and as such, these constitute the majority of experimental data in RBPDB.

A short text description of the RBDs in the largest isoform of the protein (e.g. RRMx2 for a protein with two RRM domains) was assigned, and links to UniProt were added where available. In addition, in order to facilitate comparison between the RNA-binding specificities of similar proteins in different organisms, we imported orthology relationships from InParanoid (28).

During the course of curation, when we encountered RNA-binding experiments for proteins in other species (such as Xenopus, yeast or rat), we added them to the database on an *ad hoc* basis. However, coverage of the RNA-binding proteomes of species other than human, mouse, Drosophila and *C. elegans* is not intended to be comprehensive.

Table 1. Current species and protein domain coverage in RBPDB

Species	Number of proteins
Human	422
Mouse	413
Fly (*Drosophila melanogaster*)	258
Worm (*Caenorhabditis elegans*)	244
RNA-binding domain	Number of proteins[a]
RNA Recognition Motif	733
CCCH zinc finger	225
K Homology	138
Like-Sm domain	81
C2H2 zinc finger	30
Ribosomal protein S1-like	32
Cold-shock domain	29
Lupus La RNA-binding domain	26
Pumilio-like repeat	23
Pseudouridine synthase and archaeosine transglycosylase (PUA domain)	21
Surp module/SWAP	19
Sterile Alpha Motif	11
YTH domain	12
PWI domain	10
THUMP domain	9
TROVE module	6

[a]Many proteins have more than one RBD.

Types and representation of RNA protein interactions

We populated RBPDB with RNA-binding data by searching PubMed with the gene names and aliases of the aforementioned RBPs, and recording any RNA-binding data found in the retrieved papers. RBPDB currently catalogs 14 types of RNA-binding experiments. These include experiments that measure binding to a single sequence and those that measure binding to many sequences in parallel, *in vivo* or *in vitro*. A description of the categories of experiments and the number of experiments in each category is given in Table 2.

Single-sequence experiments. Single-sequence experiments were included where the sequence of the bound RNA could be determined and is less than 200 nt in length. For these experiments, the full nucleotide sequence is included, unless a consensus motif rather than a unique sequence is reported. The consensus sequences use IUPAC (International Union of Pure and Applied Chemistry) nomenclature for representing degenerate nucleotides. Additionally, sequences with variable-length stretches or repetitive motifs are reported as (M)(X), where M is the repeated nucleotide or sequence, and X is a numerical value/range or a long undefined sequence (denoted as ‘n’). For example, the motif CUCUCU(A)(15–30)CUCU CU described for PTB contains two CUCUCU sequences separated by 15–30 adenosines (29), while (G)(n) denotes a poly(G) sequence.

SELEX experiments. For SELEX experiments, we extracted the selected sequences from the publication and aligned them as reported. We then created a position frequency matrix (PFM) from the alignment, and calculated a PWM using the Transcription Factor Binding Site (TFBS) package (30). Logos were created using the WebLogo standalone package (31). Reported motifs that contained internal gaps that would preclude representation in matrix format, or those for which >10% of the selected sequences do not match the reported motif, are reported as an IUPAC consensus motif only, as described above.

Large-scale in vivo binding experiments. When possible, we compiled all sequences identified in large-scale *in vivo* binding experiments. There is considerable diversity in how these data and sequences are reported and annotated. In some cases, we were unable to recover sequences; in these cases, RBPDB refers to the original publication but does not contain the sequences. When we were able to recover bound sequences, we included a short README file to describe how the sequences were extracted from supplementary data or GEO (Gene Expression Omnibus) (32). In general, when bound sequences were detected by tiling arrays, we extracted genomic sequence from the sense strand with respect to the annotated gene located ±200 bp of all reported peaks, since it is possible that pre-mRNA is bound, along with

Table 2. Types and numbers of experiments currently contained in RBPDB

Experiment type	Description	Number of experiments in RBPDB
EMSA	Electromobility shift assays measure binding to a single RNA sequence *in vitro* by observing a change in RNA migration rate caused by binding to protein.	522
UV cross-linking	A single radiolabeled RNA sequence is cross-linked in cellular extract using UV radiation, and the bound proteins are separated by gel electrophoresis. Protein identity is determined using mass spectrometry or a protein-specific antibody.	234
Protein affinity purification	A synthetic RNA oligo or *in vitro* transcribed RNA is derivatized with a functional group, usually biotin, which allows it to be immobilized on streptavitin beads or affinity column. Cellular extract is applied, and the proteins that bind to the RNA are identified using antibodies.	156
SELEX	High-affinity binding sequences are selected from a randomized pool by several sequential rounds of binding to purified protein and PCR amplification. The resulting RNAs are cloned and sequenced, providing a set of short sequences preferred by the protein, which are analyzed for motifs, consensus sequences and structural preferences.	117
Genome-wide RNA immunoprecipitation	These methods assay for cellular RNAs bound to a protein *in vivo*, and include RIP-chip (or RIP-seq) where RNA is purified by immunoprecipitation with an antibody to the protein (41); HITS-CLIP (or CLIP-seq), where the immunoprecipitation is preceded by UV cross-linking (CLIP) (42); and PAR-CLIP where cross-linked sites are marked by an induced thymidine to cytidine transition (43). Affinity tags and RNA fragmentation are used in some cases. RNAs are detected by microarray or sequencing. A short motif can be detected in some cases, especially if the detected RNA fragments are short and numerous.	91
Filter binding assay	A single radiolabeled RNA is incubated with protein and filtered through a nitrocellulose filter. Protein-bound RNA is retained and detected.	73
Homopolymer-binding assay	The protein is typically incubated with agarose beads bound to a homoribopolymer sequence. The preference of the protein for poly(A), poly(C), poly(G) or poly(U) can be determined.	69
NMR	Nuclear magnetic resonance spectroscopy can be used to determine nucleotide-amino-acid level interactions for RBPs.	64
Fluorescence methods	This category includes several methods of measuring binding of a protein to a single fluor-tagged RNA sequence.	47
Yeast three-hybrid assay	In the yeast three-hybrid system, a modification of the yeast two-hybrid system for measuring protein–protein interactions, binding to the RNA of interest is measured by transcription of a reporter gene in yeast.	30
Yeast three-hybrid screen	The yeast-three hybrid system is applied to a library of RNA sequences in parallel.	12
Biosensor analysis	A method of detecting interactions between biomolecules using an RNA molecule coupled to a piezoelectric crystal. Binding to the protein of interest is detected by surface plasmon resonance.	10
RNAcompete	In the RNAcompete assay, a pool of RNA designed for specific sequence and structural features is incubated in excess to a GST-tagged protein. RNAs compete to bind to the protein, and the relative enrichment in the pulldown versus the pool is determined by microarray (44).	9
Other	This category includes rare methods such as isothermal titration calorimetry, single RNA immunoprecipitation or affinity purification and enzymatic RNA footprinting.	13

any numerical value associated with the peak (e.g. log ratio intensity). When only the identity of bound genes or transcripts is reported, we compiled the transcript or gene sequence retrieved from GenBank using BioPerl (33), or from batch download files from FlyBase, and reported this sequence along with its associated numerical value. There were a variety of different normalization and reporting strategies reported in these studies, and wherever possible, we report only normalized data rather than raw data, but we capture any associated GEO or ArrayExpress (34) identifiers to allow users to access the data directly. When there are multiple samples or controls, we report each separately. In some cases, matrices or sequence logos were reported for genome wide *in vivo* immunoprecipitation experiments, and are included in the database.

Representation of RNA structural requirements

RBDs recognize specific RNA sequences, structures or both. RNA binding *in vivo* is presumably dependent on a combination of factors, including accessibility of the binding site (35) and interactions with cofactors (including other RBPs). A goal of RBPDB is to describe bound sequences with minimal interpretation, which conflicts with complications surrounding the representation and storage of RNA structure in a compact, unambiguous, computer- and human-readable format. For example, minimum free energy structures require a windowing function to select the region of RNA to fold and are too simple to represent suboptimal structures, which can be biologically functional. Therefore, in RBPDB we include only a yes/no indication of whether the original manuscripts discussed the secondary structure of the RNA. Users interested in predicting structure should consider the RNAfold webserver (among others) (36).

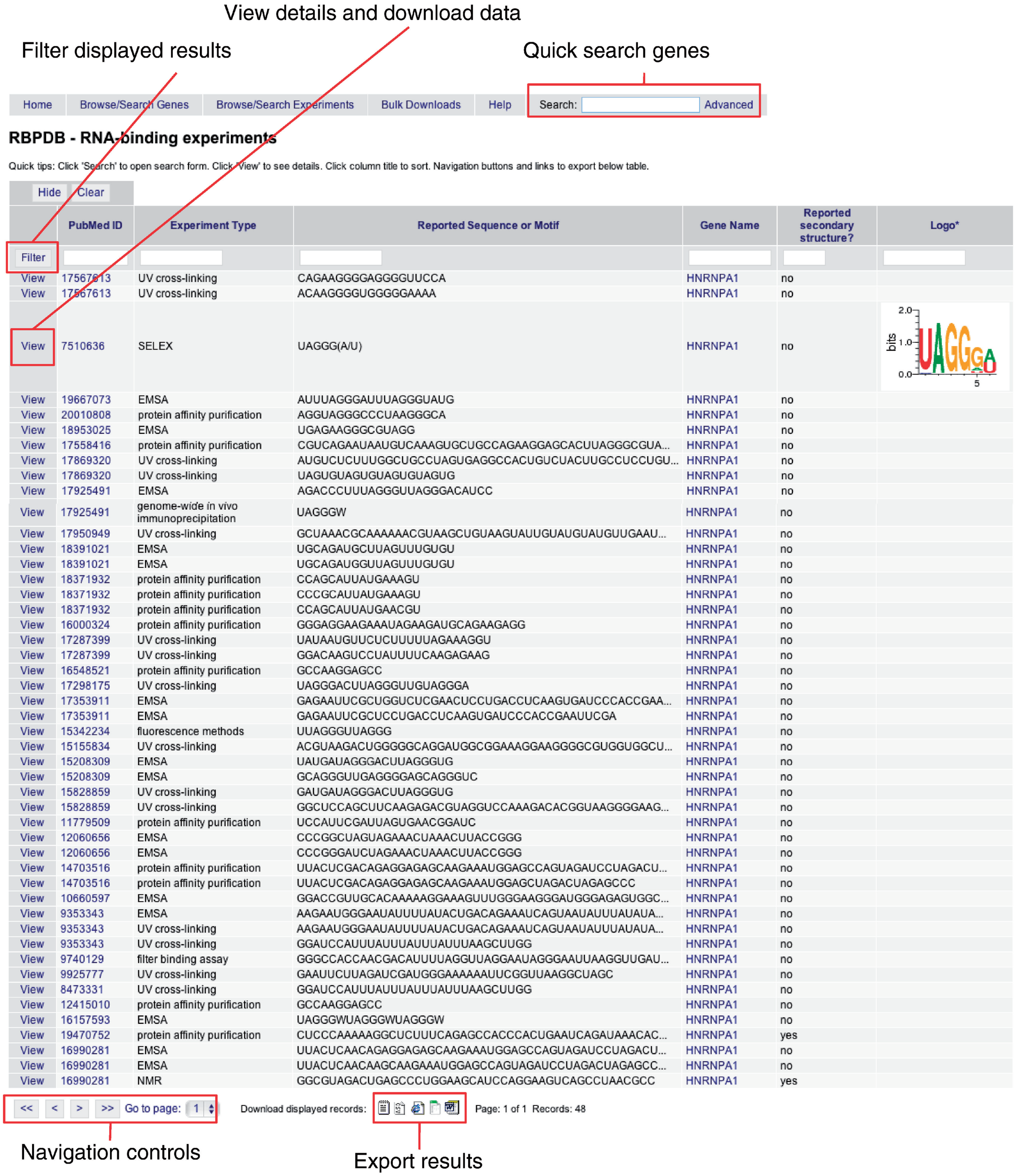
USING RBPDB

There are three main modes of interaction with RBPDB. The first is to search for RNA-binding experiments by RBP, by RBD, by species, by experiment type or by any combination of the above. The second is to perform bulk downloads of all RBPDB data or subsets of the data filtered in various ways. The third is to scan an input

RNA sequence for potential binding sites for RBPs stored in RBPDB.

Searching for RNA-binding experiments

RBPDB can be searched quickly by gene name, alias or description, by entering a search term in the search box on the home page or at the top of every page. More complex queries can be executed using the advanced search form, reached by clicking the 'advanced' link. From here, the proteins database can be searched by gene name or symbol, organism, or RBDs by making the appropriate selections on the form. To retrieve experiment records directly, the experiments form should be used; it takes



RBPDB - RNA-binding experiments

Quick tips: Click 'Search' to open search form. Click 'View' to see details. Click column title to sort. Navigation buttons and links to export below table.

	PubMed ID	Experiment Type	Reported Sequence or Motif	Gene Name	Reported secondary structure?	Logo*
View	17567613	UV cross-linking	CAGAAGGGGAGGGGUUCCA	HNRNPA1	no	
View	17567613	UV cross-linking	ACAAGGGGUGGGGGAAAA	HNRNPA1	no	
View	7510636	SELEX	UAGGG(A/U)	HNRNPA1	no	
View	19667073	EMSA	AUUUAGGGAUUUAGGGUAUG	HNRNPA1	no	
View	20010808	protein affinity purification	AGGUAGGGCCCUAAGGGCA	HNRNPA1	no	
View	18953025	EMSA	UGAGAAGGGCGUAGG	HNRNPA1	no	
View	17558416	protein affinity purification	CGUCAGAAUAAUGUCAAAGUGCUGCCAGAAGGAGCACUUAGGGCGUA...	HNRNPA1	no	
View	17869320	UV cross-linking	AUGUCUCUUUGGCUGCCUAGUGAGGCCACUGUCUACUUGCCUCCUGU...	HNRNPA1	no	
View	17869320	UV cross-linking	UAGUGUAGUGUAGUGUAGUG	HNRNPA1	no	
View	17925491	EMSA	AGACCCUUUAGGGUUAGGGACAUCC	HNRNPA1	no	
View	17925491	genome-wide in vivo immunoprecipitation	UAGGGW	HNRNPA1	no	
View	17950949	UV cross-linking	GCUAAACGCAAAAAACGUAAGCUGUAAGUAUUGUAUGUAUGUUGAAU...	HNRNPA1	no	
View	18391021	EMSA	UGCAGAUGCUUAGUUUGUGU	HNRNPA1	no	
View	18391021	EMSA	UGCAGAUGGUUAGUUUGUGU	HNRNPA1	no	
View	18371932	protein affinity purification	CCAGCAUUAUGAAAGU	HNRNPA1	no	
View	18371932	protein affinity purification	CCCGCAUUAUGAAAGU	HNRNPA1	no	
View	18371932	protein affinity purification	CCAGCAUUAUGAACGU	HNRNPA1	no	
View	16000324	protein affinity purification	GGGAGGAAGAAAUAGAAGAUGCAGAAGAGG	HNRNPA1	no	
View	17287399	UV cross-linking	UAUAAUGUUCUCUUUUUAGAAAGGU	HNRNPA1	no	
View	17287399	UV cross-linking	GGACAAGUCCUAUUUUCAAGAGAAG	HNRNPA1	no	
View	16548521	protein affinity purification	GCCAAGGAGCC	HNRNPA1	no	
View	17298175	UV cross-linking	UAGGGACUUAGGGUUGUAGGGA	HNRNPA1	no	
View	17353911	EMSA	GAGAAUUCGCUGGUCUCGAACUCCUGACCUCAAGUGAUCCCACCGAA...	HNRNPA1	no	
View	17353911	EMSA	GAGAAUUCGCUCCUGACCUCAAGUGAUCCCACCGAAUUCGA	HNRNPA1	no	
View	15342234	fluorescence methods	UUAGGGUUAGGG	HNRNPA1	no	
View	15155834	UV cross-linking	ACGUAAGACUGGGGGCAGGAUGGCGGAAAGGAAGGGGCGUGGUGGCU...	HNRNPA1	no	
View	15208309	EMSA	UAUGAUAGGGACUUAGGGUG	HNRNPA1	no	
View	15208309	EMSA	GCAGGGUUGAGGGGAGCAGGGUC	HNRNPA1	no	
View	15828859	UV cross-linking	GAUGAUAGGGACUUAGGGUG	HNRNPA1	no	
View	15828859	UV cross-linking	GGCUCCAGCUUCAAGAGACGUAGGUCCAAAGACACGGUAAGGGGAAG...	HNRNPA1	no	
View	11779509	protein affinity purification	UCCAUUCGAUUAGUGAACGGAUC	HNRNPA1	no	
View	12060656	EMSA	CCCGGCUAGUAGAAACUAAACUUACCGGG	HNRNPA1	no	
View	12060656	EMSA	CCCGGGAUCUAGAAACUAAACUUACCGGG	HNRNPA1	no	
View	14703516	protein affinity purification	UUACUCGACAGAGGAGAGCAAGAAAUGGAGCCAGUAGAUCCUAGACU...	HNRNPA1	no	
View	14703516	protein affinity purification	UUACUCGACAGAGGAGAGCAAGAAAUGGAGCUAGACUAGAGCCC	HNRNPA1	no	
View	10660597	EMSA	GGACCGUUGCACAAAAAGGAAAGUUUGGGAAGGGAUGGGAGAGUGGC...	HNRNPA1	no	
View	9353343	EMSA	AAGAAUGGGAAUAUUUUAUACUGACAGAAAUCAGUAAUAUUUAUAUA...	HNRNPA1	no	
View	9353343	UV cross-linking	AAGAAUGGGAAUAUUUUAUACUGACAGAAAUCAGUAAUAUUUAUAUA...	HNRNPA1	no	
View	9353343	UV cross-linking	GGAUCCAUUUAUUUAUUUAUUUAAGCUUGG	HNRNPA1	no	
View	9740129	filter binding assay	GGGCCACCAACGACAUUUUAGGUUAGGAAUAGGGAAUUAAGGUUGAU...	HNRNPA1	no	
View	9925777	UV cross-linking	GAAUUCUUAGAUCGAUGGGAAAAAAUUCGGUUAAGGCUAGC	HNRNPA1	no	
View	8473331	UV cross-linking	GGAUCCAUUUAUUUAUUUAUUUAAGCUUGG	HNRNPA1	no	
View	12415010	protein affinity purification	GCCAAGGAGCC	HNRNPA1	no	
View	16157593	EMSA	UAGGGWUAGGGWUAGGGW	HNRNPA1	no	
View	19470752	protein affinity purification	CUCCCAAAAAGGCUCUUUCAGAGCCACCCACUGAAUCAGAUAAACAC...	HNRNPA1	yes	
View	16990281	EMSA	UUACUCAACAGAGGAGAGCAAGAAAUGGAGCCAGUAGAUCCUAGACU...	HNRNPA1	no	
View	16990281	EMSA	UUACUCAACAAGCAAGAAAUGGAGCCAGUAGAUCCUAGACUAGAGCC...	HNRNPA1	no	
View	16990281	NMR	GGCGUAGACUGAGCCCUGGAAGCAUCCAGGAAGUCAGCCUAACGCC	HNRNPA1	yes	

Download displayed records:

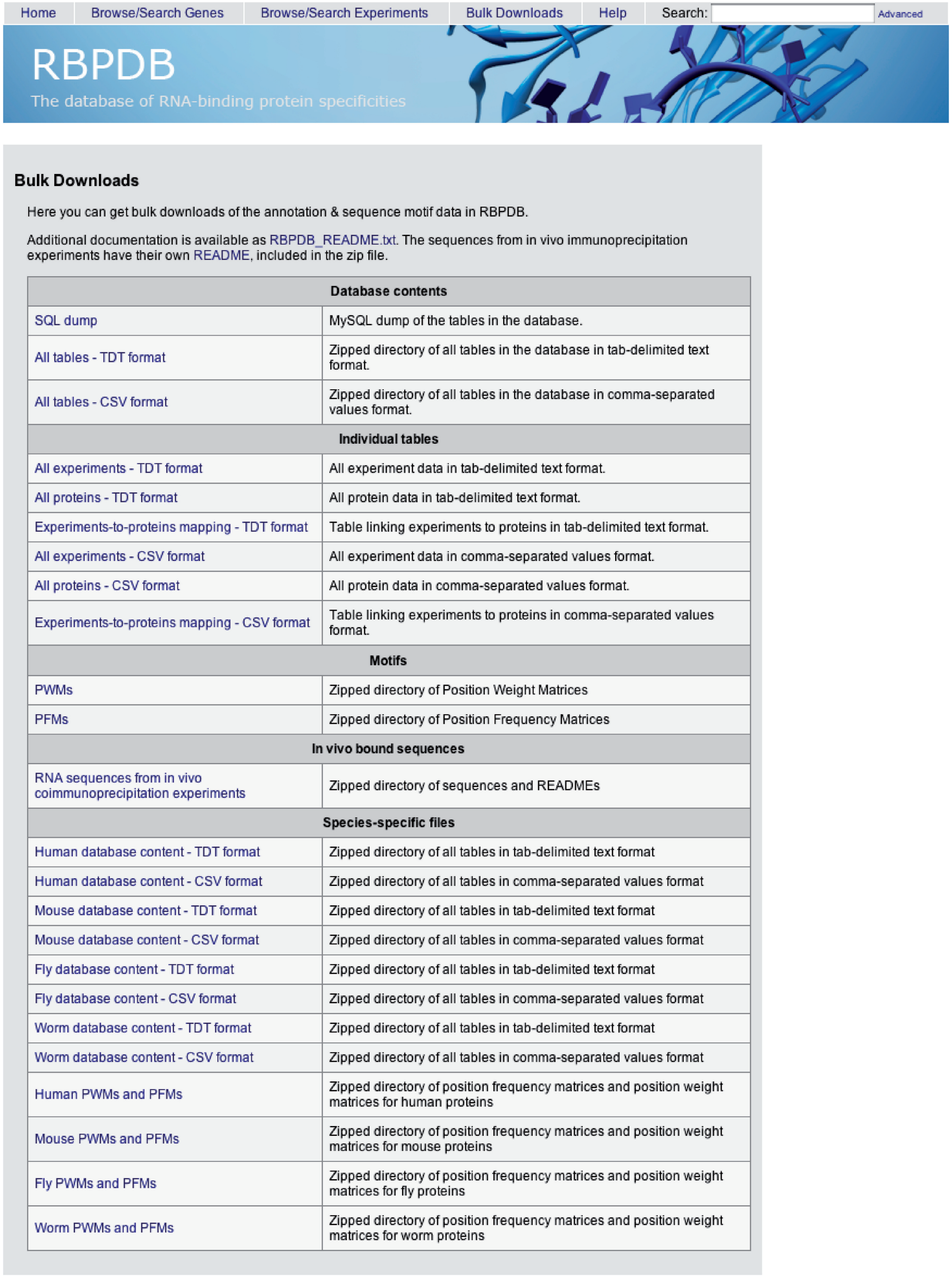
Page: 1 of 1 Records: 48

Figure 1. Example of searching RBPDB by gene name. Shown are results generated by using the advanced search form to search experiments. The query 'HNRNPA1' was entered in the gene name field and 'human' selected for species. Navigation links and links to view detailed information are indicated, as are the icons to export data in text, CSV, Excel, HTML and Word formats.

the same input, with the addition of options to search by experiment type. Figure 1 shows the results from one such search. From the results page, experimental data can be viewed and exported. Any results table can also be further filtered by partial text matches in any of the columns by clicking 'Filter'. Columns can be sorted in decreasing or increasing order by clicking the column label.

Bulk download of annotation, transcript and matrix data

There are two ways to download data from RBPDB. First, the annotation data corresponding to a subset of proteins or experiments resulting from a search query can be exported in plain text, comma separated values (CSV), Excel or Word formats directly from a search


Home
Browse/Search Genes
Browse/Search Experiments
Bulk Downloads
Help
Search:
Advanced
RBPDB
The database of RNA-binding protein specificities

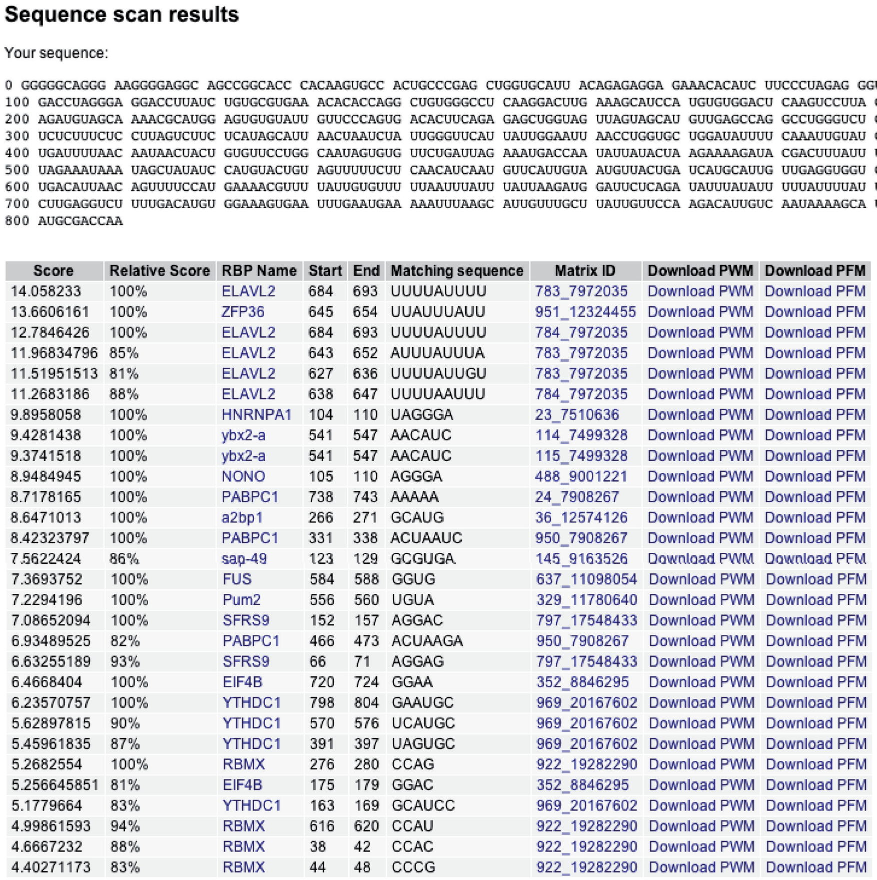
Bulk Downloads

Here you can get bulk downloads of the annotation & sequence motif data in RBPDB.

Additional documentation is available as RBPDB_README.txt. The sequences from in vivo immunoprecipitation experiments have their own README, included in the zip file.

Database contents	
SQL dump	MySQL dump of the tables in the database.
All tables - TDT format	Zipped directory of all tables in the database in tab-delimited text format.
All tables - CSV format	Zipped directory of all tables in the database in comma-separated values format.
Individual tables	
All experiments - TDT format	All experiment data in tab-delimited text format.
All proteins - TDT format	All protein data in tab-delimited text format.
Experiments-to-proteins mapping - TDT format	Table linking experiments to proteins in tab-delimited text format.
All experiments - CSV format	All experiment data in comma-separated values format.
All proteins - CSV format	All protein data in comma-separated values format.
Experiments-to-proteins mapping - CSV format	Table linking experiments to proteins in comma-separated values format.
Motifs	
PWMs	Zipped directory of Position Weight Matrices
PFMs	Zipped directory of Position Frequency Matrices
In vivo bound sequences	
RNA sequences from in vivo coimmunoprecipitation experiments	Zipped directory of sequences and READMEs
Species-specific files	
Human database content - TDT format	Zipped directory of all tables in tab-delimited text format
Human database content - CSV format	Zipped directory of all tables in comma-separated values format
Mouse database content - TDT format	Zipped directory of all tables in tab-delimited text format
Mouse database content - CSV format	Zipped directory of all tables in comma-separated values format
Fly database content - TDT format	Zipped directory of all tables in tab-delimited text format
Fly database content - CSV format	Zipped directory of all tables in comma-separated values format
Worm database content - TDT format	Zipped directory of all tables in tab-delimited text format
Worm database content - CSV format	Zipped directory of all tables in comma-separated values format
Human PWMs and PFMs	Zipped directory of position frequency matrices and position weight matrices for human proteins
Mouse PWMs and PFMs	Zipped directory of position frequency matrices and position weight matrices for mouse proteins
Fly PWMs and PFMs	Zipped directory of position frequency matrices and position weight matrices for fly proteins
Worm PWMs and PFMs	Zipped directory of position frequency matrices and position weight matrices for worm proteins

Figure 2. Download page of RBPDB. This screenshot shows the bulk data set downloads available.



Sequence scan results

Your sequence:

```
0 GGGGGCAGGG AAGGGGAGGC AGCCGGCACC CACAAGUGCC ACUGCCCGAG CUGGUGCAUU ACAGAGAGGA GAAACACAUC UUCCCUAGAG GGUUCCUGUA
100 GACCUAGGGA GGACCUUAUC UGUGCGUGAA ACACACCAGG CUGUGGGCCU CAAGGACUUG AAAGCAUCCA UGUGUGGACU CAAGUCCUUA CCUCUUCCGG
200 AGAUGUAGCA AAACGCAUGG AGUGUGUAUU GUUCCCAGUG ACACUUCAGA GAGCUGGUAG UUAGUAGCAU GUUGAGCCAG GCCUGGGUCU GUGUCUCUUU
300 UCUCUUUCUC CUUAGUCUUC UCAUAGCAUU AACUAAUCUA UUGGGUUCAU UAUUGGAAUU AACCUGGUGC UGGAUAUUUU CAAAUUGUAU CUAGUGCAGC
400 UGAUUUUAAC AAUAACUACU GUGUUCCUGG CAAUAGUGUG UUCUGAUUAG AAAUGACCAA UAUUAUACUA AGAAAAGAUA CGACUUUAUU UUCUGGUAGA
500 UAGAAAUAAA UAGCUAUAUC CAUGUACUGU AGUUUUUCUU CAACAUCAAU GUUCAUUGUA AUGUUACUGA UCAUGCAUUG UUGAGGUGGU CUGAAUGUUC
600 UGACAUUAAC AGUUUUCCAU GAAAACGUUU UAUUGUGUUU UUAAUUUAUU UAUUAAGAUG GAUUCUCAGA UAUUUAUAUU UUUAUUUUAU UUUUUUCUAC
700 CUUGAGGUCU UUUGACAUGU GGAAAGUGAA UUUGAAUGAA AAAUUUAAGC AUUGUUUGCU UAUUGUUCCA AGACAUUGUC AAUAAAAGCA UUUAAGUUGA
800 AUGCGACCAA
```

Score	Relative Score	RBP Name	Start	End	Matching sequence	Matrix ID	Download PWM	Download PFM
14.058233	100%	ELAVL2	684	693	UUUUAUUUU	783_7972035	Download PWM	Download PFM
13.6606161	100%	ZFP36	645	654	UUAUUUAUU	951_12324455	Download PWM	Download PFM
12.7846426	100%	ELAVL2	684	693	UUUUAUUUU	784_7972035	Download PWM	Download PFM
11.96834796	85%	ELAVL2	643	652	AUUUAUUUA	783_7972035	Download PWM	Download PFM
11.51951513	81%	ELAVL2	627	636	UUUUAUUGU	783_7972035	Download PWM	Download PFM
11.2683186	88%	ELAVL2	638	647	UUUUAAUUU	784_7972035	Download PWM	Download PFM
9.8958058	100%	HNRNPA1	104	110	UAGGGA	23_7510636	Download PWM	Download PFM
9.4281438	100%	ybx2-a	541	547	AACAUC	114_7499328	Download PWM	Download PFM
9.3741518	100%	ybx2-a	541	547	AACAUC	115_7499328	Download PWM	Download PFM
8.9484945	100%	NONO	105	110	AGGGA	488_9001221	Download PWM	Download PFM
8.7178165	100%	PABPC1	738	743	AAAAA	24_7908267	Download PWM	Download PFM
8.6471013	100%	a2bp1	266	271	GCAUG	36_12574126	Download PWM	Download PFM
8.42323797	100%	PABPC1	331	338	ACUAAUC	950_7908267	Download PWM	Download PFM
7.5622424	86%	sap-49	123	129	GCGUGA	145_9163526	Download PWM	Download PFM
7.3693752	100%	FUS	584	588	GGUG	637_11098054	Download PWM	Download PFM
7.2294196	100%	Pum2	556	560	UGUA	329_11780640	Download PWM	Download PFM
7.08652094	100%	SFRS9	152	157	AGGAC	797_17548433	Download PWM	Download PFM
6.93489525	82%	PABPC1	466	473	ACUAAGA	950_7908267	Download PWM	Download PFM
6.63255189	93%	SFRS9	66	71	AGGAG	797_17548433	Download PWM	Download PFM
6.4668404	100%	EIF4B	720	724	GGAA	352_8846295	Download PWM	Download PFM
6.23570757	100%	YTHDC1	798	804	GAAUGC	969_20167602	Download PWM	Download PFM
5.62897815	90%	YTHDC1	570	576	UCAUGC	969_20167602	Download PWM	Download PFM
5.45961835	87%	YTHDC1	391	397	UAGUGC	969_20167602	Download PWM	Download PFM
5.2682554	100%	RBMX	276	280	CCAG	922_19282290	Download PWM	Download PFM
5.256645851	81%	EIF4B	175	179	GGAC	352_8846295	Download PWM	Download PFM
5.1779664	83%	YTHDC1	163	169	GCAUCC	969_20167602	Download PWM	Download PFM
4.99861593	94%	RBMX	616	620	CCAU	922_19282290	Download PWM	Download PFM
4.6667232	88%	RBMX	38	42	CCAC	922_19282290	Download PWM	Download PFM
4.40271173	83%	RBMX	44	48	CCCG	922_19282290	Download PWM	Download PFM

Figure 3. Example of scanning input sequence for potential RBP-binding sites. The 3′-UTR of human c-fos was downloaded from GENBANK (Accession no. NM_005252, nucleotides 1349–2158) and submitted to the sequence scan form on the RBPDB home page.

result table, as shown in Figure 1. The second way to download data is via the Downloads page, linked from the menu at the top of the site (Figure 2). This page has links to files that include the full annotation database in SQL, tab-delimited and CSV formats, as well as sets of transcripts bound in genome wide *in vivo* experiments, and binding specificity PFM and PWM matrices in a flat text file format (30). The individual protein and experiment tables are also available, as well as the linker table needed to map experiments to proteins. These files are also available for each species separately.

Scanning input sequences for RBP-binding sites

From the main page, users can submit nucleotide sequences to scan for matches with RBP-binding sites. This sequence can be in DNA or RNA format. Additionally, a threshold for reporting matches to the sequence can be set. At present, the sequence can only be scanned with motifs associated with full PWMs. Potential binding sites in the sequence are identified by scoring potential binding sites within the sequence using PWMs, using BioPerl (33). The PWM score for a potential binding site is the sum of the scores of each nucleotide at each position in the PWM, and the relative score is the percent of the score relative to the maximum possible score of the PWM calculated. Sites with relative scores greater than the threshold, which defaults to 80%, are reported. Figure 3 shows the results obtained for the 3′-UTR of the human c-fos gene. The RBPs TTP and members of the ELAV family have been implicated in the ARE-regulated degradation of c-fos RNA (37). The top hits are to known AU-rich element (ARE)-binding proteins ELAVL2 (HuB) and ZFP36 (TTP).

It is also possible to search all individual RNA sequences from the single-sequence experiments by entering a sequence or IUPAC consensus of interest in the search window. The search will return exact matches to the text entered.

FUTURE DIRECTIONS

We will periodically update RBPDB to keep it current. Each protein entry in our database will be reassessed at least once a year. RBPDB also has a user submission form that allows users to notify our curators of recent publications of RNA-binding specificities or proteins newly discovered; we will prioritize these submissions for updates. Newly-described RBDs [e.g. the nudix domain (38)] and newly described RBPs without conserved domains will be included using the search strategy used for the initial construction of the database. A related future direction for RBPDB will be the systematic incorporation of data from other species. RBPDB is currently populated only with data from metazoans, which are of special interest for biomedical research, but represent only a small minority of the eukaryotic kingdom. There is RNA-binding information for proteins in other species, particularly traditional non-metazoan model systems such as yeast (39) and Arabidopsis [e.g. (40)], and also bacteria.

It may also be possible to further populate the database by inferring RNA-binding activities. While the existence of a universal molecular 'code' that predicts RNA sequence specificity directly from protein sequence has proven difficult to derive (25), there is little question that proteins with very similar amino-acid sequences tend to have very similar RNA-binding activities. As such, we anticipate that one application of RBPDB will be further analysis of the relationships between protein sequences and RNA-binding activities. For these analyses, it would be invaluable for the RNA-binding activities of individual RBDs to be documented, rather than individual proteins and the bound sequences to be aligned, if possible. Indeed, the way the RNA-binding activity is represented is critical for many uses of RBPDB, including genome scanning, identification of proteins that would bind sequences of interest, and comparisons among RBPs. Therefore, an area of ongoing exploration will be the representation of RNA-binding activities, including the inclusion of domain-specific information and incorporation of RNA structure.

ACKNOWLEDGEMENTS

The authors are grateful to Harm van Bakel, Debashish Ray and Carl de Boer for computational support and helpful conversations.

FUNDING

Canadian Institutes of Health Research (MOP-93671 to T.R.H. and Q.M.; MOP-49451 to T.R.H.); National Institutes of Health (1R01HG00570 to T.R.H.); Natural Sciences and Engineering Research Council of Canada CGS-M (to K.C.). Funding for open access charge: Canadian Institutes of Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Licatalosi,D.D. and Darnell,R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.
2. McKee,A.E. and Silver,P.A. (2007) Systems perspectives on mRNA processing. *Cell Res.*, **17**, 581–590.
3. Sanchez-Diaz,P. and Penalva,L.O. (2006) Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. *RNA Biol.*, **3**, 101–109.
4. Dreyfuss,G., Kim,V.N. and Kataoka,N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.*, **3**, 195–205.
5. Rodriguez,A.J., Czaplinski,K., Condeelis,J.S. and Singer,R.H. (2008) Mechanisms and cellular roles of local protein synthesis in mammalian cells. *Curr. Opin. Cell Biol.*, **20**, 144–149.
6. Blencowe,B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37–47.
7. Anantharaman,V., Koonin,E.V. and Aravind,L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.*, **30**, 1427–1464.
8. Clery,A., Blatter,M. and Allain,F.H. (2008) RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.*, **18**, 290–298.
9. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
10. Oberstrass,F.C., Auweter,S.D., Erat,M., Hargous,Y., Henning,A., Wenter,P., Reymond,L., Amir-Ahmady,B., Pitsch,S., Black,D.L. *et al.* (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, **309**, 2054–2057.
11. Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Eppig,J.T. and Mouse Genome Database,G. (2010) The Mouse Genome Database: enhancements and updates. *Nucleic Acids Res.*, **38**, D586–D592.
12. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
13. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
14. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
15. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
18. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin,J.A., Eppig,J.T., Baldarelli,R.M., Barsanti,K., Baya,M., Beal,J.S., Boddy,W.J. *et al.* (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, D476–D481.
19. Achsel,T., Stark,H. and Luhrmann,R. (2001) The Sm domain is an ancient RNA-binding motif with oligo(U) specificity. *Proc. Natl Acad. Sci. USA*, **98**, 3685–3689.
20. Worbs,M., Bourenkov,G.P., Bartunik,H.D., Huber,R. and Wahl,M.C. (2001) An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol. Cell*, **7**, 1177–1189.
21. Hall,T.M. (2005) Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.*, **15**, 367–373.
22. Denhez,F. and Lafyatis,R. (1994) Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot. *J. Biol. Chem.*, **269**, 16170–16179.
23. Aravind,L. and Koonin,E.V. (2001) THUMP–a predicted RNA-binding domain shared by 4-thiouridine, pseudouridine synthases and RNA methylases. *Trends Biochem. Sci.*, **26**, 215–217.
24. Aviv,T., Lin,Z., Lau,S., Rendl,L.M., Sicheri,F. and Smibert,C.A. (2003) The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. *Nat. Struct. Biol.*, **10**, 614–621.
25. Auweter,S.D., Oberstrass,F.C. and Allain,F.H. (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.*, **34**, 4943–4959.
26. Szymczyna,B.R., Bowman,J., McCracken,S., Pineda-Lucena,A., Lu,Y., Cox,B., Lambermon,M., Graveley,B.R., Arrowsmith,C.H. and Blencowe,B.J. (2003) Structure and function of the PWI

motif: a novel nucleic acid-binding domain that facilitates pre-mRNA processing. *Genes Dev*., **17**, 461–475.

27. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
28. Berglund,A.C., Sjolund,E., Ostlund,G. and Sonnhammer,E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*., **36**, D263–D266.
29. Lamichhane,R., Daubner,G.M., Thomas-Crusells,J., Auweter,S.D., Manatschal,C., Austin,K.S., Valniuk,O., Allain,F.H. and Rueda,D. (2010) RNA looping by PTB: evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc. Natl Acad. Sci. USA*, **107**, 4105–4110.
30. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
31. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.
32. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*., **35**, D760–D765.
33. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
34. Kapushesky,M., Emam,I., Holloway,E., Kurnosov,P., Zorin,A., Malone,J., Rustici,G., Williams,E., Parkinson,H. and Brazma,A. (2010) Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res*., **38**, D690–D698.
35. Li,X., Quon,G., Lipshitz,H.D. and Morris,Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.
36. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res*., **36**, W70–W74.
37. Chen,C.Y., Gherzi,R., Ong,S.E., Chan,E.L., Raijmakers,R., Pruijn,G.J., Stoecklin,G., Moroni,C., Mann,M. and Karin,M. (2001) AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell*, **107**, 451–464.
38. Yang,Q., Gilmartin,G.M. and Doublie,S. (2010) Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3′ processing. *Proc. Natl Acad. Sci. USA*, **107**, 10062–10067.
39. Hogan,D.J., Riordan,D.P., Gerber,A.P., Herschlag,D. and Brown,P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*., **6**, e255.
40. Tam,P.P., Barrette-Ng,I.H., Simon,D.M., Tam,M.W., Ang,A.L. and Muench,D.G. (2010) The Puf family of RNA-binding proteins in plants: phylogeny, structural modeling, activity and subcellular localization. *BMC Plant Biol*., **10**, 44.
41. Tenenbaum,S.A., Carson,C.C., Lager,P.J. and Keene,J.D. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl Acad. Sci. USA*, **97**, 14085–14090.
42. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
43. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
44. Ray,D., Kazan,H., Chan,E.T., Castillo,L.P., Chaudhry,S., Talukder,S., Blencowe,B.J., Morris,Q. and Hughes,T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol*., **27**, 667–670.